DOCUMENT RESUME

ED 408 304 TM 026 506

AUTHOR Kim, Seock-Ho; Cohen, Allan S.
TITLE An Investigation of the Likelihood Ratio Test for Detection of Differential Item Functioning under the Graded Response Model.
PUB DATE Mar 97
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; Classification; Computer Simulation; Estimation (Mathematics); Identification; *Item Bias; *Maximum Likelihood Statistics; Monte Carlo Methods; Sample Size; *Test Items
IDENTIFIERS *Graded Response Model; Item Bias Detection; *Likelihood Ratio Tests; MULTILOG Computer Program; Type I Errors

ABSTRACT

Type I error rates of the likelihood ratio test for the detection of differential item functioning (DIF) were investigated using Monte Carlo simulations. The graded response model with five ordered categories was used to generate data sets of a 30-item test for samples of 300 and 1,000 simulated examinees. All DIF comparisons were simulated by randomly pairing two groups of examinees. Three different sample sizes of reference and focal groups comparisons were simulated under two different ability matching conditions. For each of the six combinations of sample sizes by ability matching conditions, 100 replications of DIF detection comparisons were simulated. Item parameter estimates and likelihood values were obtained by marginal maximum likelihood estimation using the computer program MULTILOG. Type I error rates of the likelihood ratio test statistics for all six combinations of the sample sizes and ability matching conditions were within theoretically expected values at each of the nominal alpha levels considered. (Contains 5 figures, 5 tables, and 24 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 408 304

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
SEOCK-HO KIM
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

An Investigation of the Likelihood Ratio Test for Detection of Differential Item Functioning Under the Graded Response Model

Seock-Ho Kim
The University of Georgia
Allan S. Cohen
University of Wisconsin–Madison

March, 1997
Running Head: LIKELIHOOD RATIO TEST UNDER THE GRADED RESPONSE MODEL

Paper presented at the annual meeting of the American Educational Research Association, Chicago.

BEST COPY AVAILABLE

TM026506

An Investigation of the Likelihood Ratio Test for Detection of Differential Item Functioning Under the Graded Response Model

Abstract

Type I error rates of the likelihood ratio test for the detection of differential item functioning (DIF) were investigated using Monte Carlo simulations. The graded response model with five ordered categories was used to generate data sets of a 30-item test for samples of 300 and 1,000 simulated examinees. All DIF comparisons were simulated by randomly pairing two groups of examinees. Three different sample sizes of reference and focal groups comparisons were simulated under two different ability matching conditions. For each of the six combinations of sample sizes by ability matching conditions, 100 replications of DIF detection comparisons were simulated. Item parameter estimates and likelihood values were obtained by marginal maximum likelihood estimation using the computer program MULTILOG. Type I error rates of the likelihood ratio test statistics for all six combinations of the sample sizes and ability matching conditions were within theoretically expected values at each of the nominal alpha levels considered.

Index terms: differential item functioning, graded response model, item response theory, likelihood ratio test, Type I error.

Introduction

In the context of dichotomously scored item response theory (IRT) models, an item is said to be functioning differentially when the probability of a correct response to the item is different for examinees at the same ability level but from different groups (Pine, 1977). In the typical differential item functioning (DIF) study, there are two groups of examinees, the reference group and the focal group. For polytomous IRT models, an item is considered to be functioning differentially when the item true score functions in the reference and focal groups are not equal (Cohen, Kim, & Baker, 1993). The presence of such items on a test is a threat to validity and also may interfere seriously with efforts to equate tests, where equating is necessary.

Thissen, Steinberg, and Gerrard (1986) and Thissen, Steinberg, and Wainer (1988, 1993) proposed the likelihood ratio test (Neyman & Pearson, 1928) to evaluate the significance of observed differences in item responses from different groups under IRT. Using a dichotomous IRT model, Kim and Cohen (1995) compared this likelihood ratio test with Lord's (1980) chi-square test, and Raju's area measures (1988, 1990), and found them to provide comparable results. Cohen, Kim, and Wollack (1996) subsequently reported Type I error rates of the likelihood ratio test for DIF under the two- and three-parameter IRT models to be within expected limits at the nominal alpha levels considered. Ankenmann, Witt, and Dunbar (1996) compared the power and Type I error rates of the likelihood ratio test and the Mantel (1963) test for DIF detection under the graded response model. Ankenmann et al. (1996) used combined dichotomous and graded response item response data and obtained the power and Type I error rates for a single studied graded response item in each data set. Type I error rates were obtained for the six different types of studied items under different sample sizes and ability conditions. The likelihood ratio test was found to yield better power and control of Type I error than the Mantel procedure (Ankenmann et al., 1996).

It is important to note that investigation of the power of a test statistic is meaningless without an adequate control of the probability of a Type I error. Previous research has been suggestive but does not provide sufficient information about the Type I error control of the likelihood ratio test for DIF in the graded response model. The present study, therefore, is designed to examine the Type I error rates of the likelihood ratio test under the graded response model using a wider variety of underlying item parameters, sample sizes and ability conditions than previously reported.

The basic building block of IRT is the item response function (IRF). For a dichotomously scored item the IRF is, in fact, the same as the item true score function that describes the functional relationship between the probability of a correct response to an item and examinee trait level, θ. For a polytomously scored item, the item true score function describes the relationship between the expected value of the item score and examinee trait level. In the context of IRT for both dichotomous and polytomous IRT models, an item functions differentially if the item true score functions obtained from different groups of examinees are different. It is important to note that item true score functions can be identical if and only if the sets of item parameters estimated in different groups are equal.

For the polytomous IRT models, the equality of sets of item parameters can be tested using several different approaches. One approach is to compare item parameters estimated from two groups of examinees (e.g., Cohen, Kim, & Baker, 1993). A second approach is to compare item true score functions estimated from two groups of examinees by measuring the areas between them (e.g., Cohen, Kim, & Baker, 1993; Flowers, Oshima, & Raju, 1995). A third approach is to compare likelihood functions, using a likelihood ratio, to evaluate the differences between item responses from two groups. Thissen et al. (1988) noted that the third approach is preferable for theoretical reasons. Also, the first and second approaches may require estimates of variances and covariances of the item parameters. At the present time, computational difficulties continue to impede obtaining accurate estimates of these variances and covariances.

DIF studies under IRT require that estimates of item parameters obtained in different groups be placed on a common metric before comparisons are made (Stocking & Lord, 1983). In the first and second approaches mentioned above a common metric can usually be obtained by calibrating item parameters in different groups and then subsequently applying a method for transforming the parameter estimates onto a base metric. Generally, this metric is obtained from the reference group data. Currently there exist several such linking methods for the polytomous IRT models (e.g., Baker, 1992). Given a common metric, DIF measures can be viewed as some function of residuals leftover after linking and expressed either in terms of the discrepancy of the parameter estimates or the discrepancy of the item true score functions (N. S. Raju & T. C. Oshima, personal communications, March 23, 1995). It is further interesting to note that the procedures used in linking are nearly the same, sometimes exactly the same, as procedures used in DIF detection. Given such

close relationships between methods of detection of DIF and methods of linking (cf., Kim & Cohen, 1992), it is easy to understand why DIF detection is prone to errors due to linking (cf., Shepard, Camilli, & Williams, 1984).

For the likelihood ratio test of DIF, using the computer program MULTILOG (Thissen, 1991), such transformations or linking of metrics are unnecessary because item parameters are estimated simultaneously in a data set consisting of both the reference and focal groups combined. The problem of a common metric for the likelihood ratio test for DIF is handled through the common or anchor set of items rather than by linking. In the likelihood ratio test, the likelihood from a compact model, in which no group differences are assumed to be present, is compared to that from an augmented model, in which one or more items are examined for possible DIF. Clearly, the metric of the compact and augmented models are dependent on the anchor items. The tentative assumption in this approach is that there are no DIF items among the common items in the compact model and methods for obtaining an appropriate compact model are important.

The comparison between a compact model and an augmented model requires two separate calibration runs for obtaining the likelihoods. For the dichotomous IRT models, Thissen et al. (1993) recommended the use of the Mantel Haenszel χ^2 (MH) for identifying a set of common items for the metric establishment purpose. This approach is generally useful but also is suspect as the MH test is not sensitive to non-uniform DIF. Kim and Cohen (1995) recommended an iterative purification method for the likelihood ratio test. The iterative purification method is quite labor intensive, however, but it is theoretically more consistent with the likelihood ratio test. For the polytomous IRT models one can apply either Mantel's (1963) procedure or a similar method based on Kim and Cohen (1995).

In marginal maximum likelihood estimation under MULTILOG, the population ability distribution may play an important role in metric construction, especially when two groups of examinees are not comparable in terms of their underlying ability. When the item response data from the reference and focal groups are combined, the item parameters are calibrated from the marginalized likelihood function over the ability parameters. Using the likelihood ratio for DIF detection via MULTILOG, the default options assume that the reference group ability parameters are normally distributed with mean 0 and variance 1. In addition, the mean of the focal group ability parameters is typically estimated in the calibration run along with item parameters, while the variance of the population ability is fixed to be 1. The effect

of this assumption, however, is not clear. In this study, therefore, in addition to the Type I error rates of the likelihood ratio test statistic, the estimated values of the focal group population mean will be compared to the theoretical values.

Method

Data Generation

In a typical DIF study, there are two groups of examinees, the reference group and the focal group. The reference group is considered to be the base group against which the parameters estimated in the focal group are compared. For the reference and focal groups in this study, two sample sizes were used to simulate small sample ($N = 300$) and large sample ($N = 1,000$) conditions. Three different sample size combinations of reference and focal groups were simulated: (1) a reference group with 300 examinees and a focal group with 300 examinees (R300/F300), (2) a reference group with 1,000 examinees and a focal group with 1,000 examinees (R1000/F1000), and (3) a reference group with 1,000 examinees and a focal group with 300 examinees (R1000/F300). The large and small sample size comparisons were selected based on previous recovery study results for the graded response model by Reise and Yu (1990) which indicated that at least 500 examinees were needed to achieve an adequate calibration.

The computer program GENIRV (Baker, 1988) was used to generate graded response model (Samejima, 1969, 1972) data sets for a 30-item test with five ordered categories. The category response function $P_{jk}(\theta)$ is the probability of response k to item j as a function of θ. For a five-category item, $P_{jk}(\theta)$ is defined as

$$P_{jk}(\theta) = \begin{cases} 1 - P^*_{j1}(\theta) & \text{when } k = 1 \\ P^*_{j(k-1)}(\theta) - P^*_{jk}(\theta) & \text{when } k = 2, 3, 4 \\ P^*_{j4}(\theta) & \text{when } k = 5. \end{cases} \tag{1}$$

In equation (1), $P^*_{jk}(\theta)$is the boundary response function ($k = 1, 2, 3, 4$) given by

$$P^*_{jk}(\theta) = \{1 + \exp[-\alpha_j(\theta - \beta_{jk})]\}^{-1}, \tag{2}$$

where α_j is the discrimination parameter for item j, β_{jk} is the location parameter, and θ is the trait level parameter. With $P^*_{j0}(\theta) = 1$ and $P^*_{j5} = 0$, the category response function can be succinctly written as

$$P_{jk}(\theta) = P_{j(k-1)}(\theta) - P^*_{jk}(\theta), \tag{3}$$

where $k = 1(1)5$.

The generating item parameters used in this study are given in Table 1. These values are based on parameter estimates from 4th, 8th and 10th grade students' responses to the mathematics assessment tests of the Wisconsin Student Assessment System (Webb, 1994). Note that the average value of the location parameters β_{jk} is .962 and the standard deviation is .893. For a group of examinees whose ability parameters are distributed as a standard normal distribution [i.e., N(0,1)], in other words, this will be a relatively difficult test. In order to match the ability of the reference group with the difficulty of the test, therefore, the ability of the reference group in this study was assumed to be normally distributed with mean 1 and standard deviation 1, that is N(1,1). In this way, the test difficulty was essentially matched to the mean ability of the reference group.

Insert Table 1 about here

For each of the three sample sizes, two different ability matching conditions were simulated: (1) an unmatched condition in which the reference group had a higher underlying ability distribution [$\theta \sim$ N(1,1)] than that of the focal group [$\theta \sim$ N(0,1)] and (2) a matched condition in which both the reference and focal groups of examinees had the same underlying ability distribution [$\theta \sim$ N(1,1)]. 100 replications were simulated for each of the six combinations of three sample sizes by two ability matching conditions.

Item Parameter Estimation

Item parameter estimates for each pair of reference and focal groups were obtained using the default options available for the marginal maximum likelihood estimation algorithm for the graded response model implemented in the computer program MULTILOG (Thissen, 1991).

The Likelihood Ratio Test

The likelihood ratio test for DIF described by Thissen et al. (1988, 1993) compares two different models—a compact model and an augmented model. The likelihood ratio statistic G^2 is the difference between the values of -2 times the log likelihood for the compact model (L_C) and -2 times the log likelihood for the augmented model (L_A). The values of the quantity -2 times the log likelihood can be obtained from the output of MULTILOG and

are based on the results over the entire data set following marginal maximum likelihood estimation. The G^2 can be written as

$$G^2 = -2\log L_C - (-2\log L_A) = -2\log L_C + 2\log L_A \quad (4)$$

and is distributed as a χ^2 under the null hypothesis with degrees of freedom equal to the difference in the number of parameters estimated in the compact and augmented models. For this study, we tested one item at a time, meaning that each G^2 was distributed as a χ^2 with 5 degrees of freedom.

In the compact model, the item parameters are assumed to be the same for both the reference and focal groups. MULTILOG has an option that permits equality constraints to be placed on items for estimation of the compact model. In this study, the parameter estimates for all 30 items in the compact model were set to be equal in both the reference and focal groups. In the augmented model, item parameters for all items except the studied item were constrained to be equal in both the reference and focal groups. These constrained items are referred to as the common or anchor set. In a DIF comparison, in other words, only the item parameters for the studied item are estimated separately in the reference and focal groups. For example, for the augmented model in which Item 1 was the studied item, item parameter estimates for Item 1 were unconstrained in the reference and focal groups. Items 2–30 formed the anchor set for this augmented model and so were each respectively constrained to have the same parameter estimates in both groups. The metric used in the likelihood ratio test, therefore, is based on the set of items contained in the anchor set. In this study, the augmented models were constructed to study a single item at a time and all items were studied sequentially for DIF.

Error Rates

Error rates were obtained by comparing the number of significant G^2s to the total number of augmented model calibration runs conducted for a given sample size and ability condition. For a single test, 31 separate calibration runs were required to estimate the necessary likelihood statistics—one run to estimate the likelihood for the compact model and 30 runs for each of the augmented models (i.e., one augmented model for each of the 30 items). For the 100 pairs of reference and focal groups in a sample size by ability matching condition, therefore, 3,100 separate calibration runs were required. For all six sample size by ability matching conditions, a total of 18,600 MULTILOG calibration runs were required.

Results

Table 1 shows the number of significant G^2s for each item at $\alpha = .05$. The data in this table illustrate the general pattern of results obtained. For Item 1, in the R300/F300 condition, for example, 9 significant G^2s were obtained for the unmatched ability condition and 3 for the matched ability condition. Since 100 replications were generated, the expected number of significant G^2s due to chance for a single item would be 5 at a nominal alpha of .05. For this same sample size condition, there were a total of 159 significant G^2 obtained across all 30 items for the unmatched ability condition and 155 for the matched ability condition. A similar pattern of results was found at all other α levels examined.

Insert Table 2 about here

Table 2 shows the number of significant G^2s for all sample size and ability conditions for different alpha levels. For example, the R300/F300 sample size at $\alpha = .05$ yielded 159 significant G^2s for the unmatched ability condition and 155 for the matched ability condition. The expected number of significant G^2s due to chance over the 100 replications at a nominal alpha level of .05 would be 150.

The bottom row of Table 2 contains the expected number of significant G^2s for the alpha levels considered in this study. Note that the observed numbers of significant G^2s at each alpha level were very close to the theoretically expected values for all the sample size by ability matching conditions.

Type I Error Rates

The results of the Type I error rates are presented in Table 3. The Type I error rates are the percentages of significant G^2s over all replications. In addition, error rates for the three sample sizes are illustrated in Figures 1a, and 1b, for the unmatched ability condition and the matched ability condition, respectively.

Insert Table 3 and Figures 1a, and 1b about here

For the unmatched ability condition, both R300/F300 and R1000/F300 yielded slightly inflated Type I error rates, especially for .05 and .1 nominal alpha levels. The R1000/F1000

DIF comparisons for the unmatched condition yielded error rates lower than the expected values for .05 and .1 nominal alpha levels.

For the matched ability condition, all sample size combinations yielded slightly inflated Type I error rates except for the very smallest nominal alpha levels considered. Type I error rates for R1000/F1000 DIF comparisons were very close to the theoretically expected values.

Insert Figures 2a, 2b, and 2c about here

Results of error rates for the two ability matching conditions are illustrated in Figures 2a, 2b and 2c, for R300/F300, R1000/F1000, and R1000/F300, respectively. For both R300/F300 and R1000/F1000 comparisons, Type I error rates were slightly closer to the theoretically expected values in the matched ability condition.

For R1000/F300 in the unmatched ability condition, the Type I error rate was slightly closer to the theoretically expected value at the .05 nominal alpha level. It is very important to note, however, that the observed differences were relatively small. Most of the Type I error rates, in fact, were quite close to the theoretically expected values in all conditions simulated.

Relationships Among Generating Parameters and Significant G^2s

Table 4 shows the Spearman rank-order correlations between the generating parameters and the number of significant G^2s at a nominal alpha level of .05. The correlations between item discrimination parameters and item location parameters were all positive except for β_{j1}. No consistent pattern of correlations was observed between generating item parameters and the numbers of significant G^2s.

Insert Table 4 about here

Estimates of the Focal Group Population Ability

Table 5 shows the average value and standard deviation of the estimates of the population parameter (i.e., $\hat{\mu}$) over 100 replications for the conditions simulated in this study. For all unmatched conditions the $\hat{\mu}$s of the focal group were approximately -1. For the matched ability conditions, the $\hat{\mu}$s were close to 0. Recall that the focal group μ values for the

generating ability parameters for the unmatched ability and the matched ability comparisons were 0 and 1, respectively. The metric of estimates from a MULTILOG calibration run was based on the reference group ability which was converted to N(0,1). The expected values of $\hat{\mu}$ for the focal group would be -1 for the unmatched condition and 0 for the matched condition. Note that the sizes of the standard deviations were very small. In fact, all cases yielded values of the focal group population mean that were very close to the expected values.

Insert Table 5 about here

Summary and Discussion

Type I error rates for G^2 for the graded response model were very close to those expected for all sample sizes at most of the alpha levels considered in both the unmatched and matched ability conditions. Results for the small sample size condition, however, did differ slightly from expected values, albeit at very small nominal alpha levels. One reasons this may have occurred is that, as Reise and Yu (1990) have suggested, samples of 300 examinees may be too small for adequate recovery of the underlying parameters under the graded response model. Even so, the Type I error rates for the small sample condition R300/F300 as well as for the R1000/F300 were generally close to the theoretically expected values.

The primary concern in most DIF studies is to be able to detect all items that function differentially. This is normally accomplished by setting the Type I error rate level high, for example, at .05 or .10, suggesting that is it is preferable to falsely identify an item as functioning differentially than to miss a true DIF item. At such alpha levels, the likelihood ratio test was found to provided good Type I error control for the sample sizes and ability matching conditions simulated. In DIF studies, however, there is also a concern for the power of the DIF statistic, that is, the extent to which it provides control over Type II errors. Such errors occur when DIF items fail to be detected. Thus far, very little work has been done on this issue. Only the work of Ankenmann et al. (1996) is available regarding the power of the likelihood ratio test. More power studies are clearly needed.

In this study, the underlying ability distribution for the reference group was set to be N(1,1). The reason for this choice was to match the distribution of ability with that of the item location parameters. There were two different underlying distributions for the focal group: N(0,1) for the unmatched condition and N(1,1) for the matched condition. No

systematic differences in the Type I error rates were observed between the unmatched and matched ability conditions.

The recovery of the underlying ability population parameter for the focal group appeared to be acceptable. In the light of the excellent Type I error control of the likelihood ratio test under different sample sizes and ability matching conditions this may not be surprising. Because the likelihood ratio test of DIF via MULTILOG does not require any metric transformation, the results of the Type I error rate do not contain errors from linking. However, the existence of possible DIF items in the anchor set is likely to affect the Type I error control and, subsequently, the power of the likelihood ratio test statistic. Additional work is needed on methods for construction of the anchor sets of items and for scale purification with the likelihood ratio test.

References

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1996, April). *An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Baker, F. B. (1988). *GENIRV: A program to generate item response vectors* [Computer program]. Madison, University of Wisconsin, Department of Educational Psychology, Laboratory of Experimental Design.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16*, 87–96.

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*, 335–350.

Cohen, A. S., Kim, S.-H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*, 15–26.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1995, April). *A Monte Carlo assessment of DFIT with polytomously scored unidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Kim, S.-H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*, 51–66.

Kim, S.-H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ration test in detection of differential item functioning. *Applied Measurement in Education, 8*, 291–312.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690–700.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I and Part II. *Biometrika, 20A*, 174–240, 263–294.

Pine, S. M. (1977). Application of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Application of computerized adaptive testing: Proceedings of a symposium presented at the 18th annual convention of the Military Testing Association* (Research Rep. No. 77-1, pp. 37–43). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197–207.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133–144.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Samejima, F. (1972). A general model for free response data. *Psychometrika Monograph Supplement*, No. 18.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*, 93–128.

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 207–210.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory* [Computer program]. Chicago: Scientific Software.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118–128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum

Webb, N. L. (1994). *Wisconsin performance assessment development project: Analysis and technical report for fiscal year 1993-94*. Madison: University of Wisconsin. Wisconsin Center for Educational Research.

Table 1

Generating Item Parameters and Number of Significant G^2s at $\alpha = .05$ for Sample Size and Ability Matching Conditions

Item	Parameters					R300/F300		R1000/F1000		R1000/F300	
	α_j	β_{j1}	β_{j2}	β_{j3}	β_{j4}	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched
1	1.46	−.35	.67	.97	1.94	9	3	4	5	4	4
2	1.73	.18	.90	1.29	1.94	7	5	4	3	6	8
3	1.81	−.37	.03	.91	2.29	6	4	4	3	2	4
4	1.53	−.56	−.13	.80	2.22	4	12	2	5	7	8
5	1.57	−.38	.49	1.04	2.33	7	8	12	8	8	5
6	1.89	−.61	.63	1.37	2.34	3	1	3	5	4	4
7	1.89	.01	.67	1.33	2.18	5	4	6	4	8	7
8	1.84	−.23	.31	.98	2.46	8	5	4	4	5	7
9	1.93	−.31	.60	1.27	2.44	7	7	9	6	7	8
10	2.53	−.36	.53	1.20	2.34	5	7	4	6	3	7
11	1.79	−.52	.39	1.54	2.00	5	4	9	6	5	4
12	1.86	−.53	−.12	1.27	2.25	4	6	4	4	7	3
13	2.35	.06	.99	1.50	2.20	5	6	5	2	3	5
14	1.79	−.20	.49	1.00	2.40	6	3	6	6	5	3
15	2.12	.20	.56	1.40	2.00	7	9	1	10	5	4
16	2.07	−.44	.18	1.34	2.15	5	6	5	7	7	6
17	2.19	−.01	.39	1.36	2.01	8	7	5	6	4	5
18	2.40	.10	1.06	1.61	2.01	5	6	7	10	12	9
19	1.79	−.10	.35	1.01	2.22	4	10	2	5	7	12
20	2.12	.19	1.10	1.45	2.01	4	2	9	6	6	7
21	1.75	−.57	.93	1.31	2.01	6	2	4	3	4	4
22	2.16	.59	.91	1.32	2.01	5	5	5	5	6	6
23	1.86	−.02	.63	1.28	2.01	5	3	8	7	4	7
24	2.22	.52	.85	1.43	2.01	5	7	4	3	10	6
25	2.18	−.27	.58	1.24	2.25	4	4	6	3	3	2
26	2.01	−.66	.41	1.63	2.24	3	5	2	4	3	3
27	2.14	.05	.71	1.03	2.09	6	5	2	5	4	7
28	2.13	.43	1.15	1.47	2.06	4	3	5	5	2	3
29	2.12	.08	.70	1.12	2.09	5	2	3	2	4	2
30	2.05	.19	.61	.94	2.38	2	4	4	5	1	5
Total						159	155	148	153	156	165

Table 2

Number of Significant G^2's for Sample Size and Ability Matching Conditions at α Levels From .0005 to .1

		α Level					
Sample Size	Ability	.0005	.001	.005	.01	.05	.1
R300/F300	Unmatched	1	2	12	29	159	317
R300/F300	Matched	5	7	17	37	155	308
R1000/F1000	Unmatched	1	3	18	33	148	291
R1000/F1000	Matched	1	1	6	30	153	301
R1000/F300	Unmatched	1	2	12	26	156	313
R1000/F300	Matched	3	6	13	34	165	303
Expected Value		1.5	3	15	30	150	300

Table 3

Proportion of Significant G^2's for Sample Size and Ability Matching Conditions at α Levels From .0005 to .1000

		α Level					
Sample Size	Ability	.0005	.0010	.0050	.0100	.0500	.1000
R300/F300	Unmatched	.0003	.0007	.0040	.0097	.0530	.1057
R300/F300	Matched	.0017	.0023	.0057	.0123	.0517	.1027
R1000/F1000	Unmatched	.0003	.0010	.0060	.0110	.0493	.0970
R1000/F1000	Matched	.0003	.0003	.0020	.0100	.0510	.1003
R1000/F300	Unmatched	.0003	.0007	.0040	.0087	.0520	.1043
R1000/F300	Matched	.0010	.0020	.0043	.0113	.0550	.1010

Table 4

Spearman ρs Among Generating Item Parameters and the Number of Significant G^2s at $\alpha = .05$

Sample Size	Ability		Generating Parameter				
			α_j	β_{j1}	β_{j2}	β_{j3}	β_{j4}
Generating Parameters		α_j	1.000				
		β_{j1}	.506	1.000			
		β_{j2}	.389	.614	1.000		
		β_{j3}	.478	.188	.434	1.000	
		β_{j4}	−.049	−.380	−.454	−.448	1.000
R300/F300	Unmatched		−.210	.069	−.027	−.215	−.218
R300/F300	Matched		.138	.011	−.416	−.047	.101
R1000/F1000	Unmatched		.068	.098	.206	.252	−.020
R1000/F1000	Matched		.032	.007	−.144	.131	−.080
R1000/F3001	Unmatched		−.177	.056	−.076	.104	−.144
R1000/F300	Matched		.002	.205	.069	−.088	−.058

Table 5

Mean and Standard Deviation of Population $\hat{\mu}$ over 100 Replications

		Population $\hat{\mu}$	
Sample Size	Ability	Mean	SD
R300/F300	Unmatched	−1.029	.024
R300/F300	Matched	.004	.018
R1000/F1000	Unmatched	−1.032	.012
R1000/F1000	Matched	.005	.009
R1000/F300	Unmatched	−1.017	.017
R1000/F300	Matched	.004	.013

Figure Captions

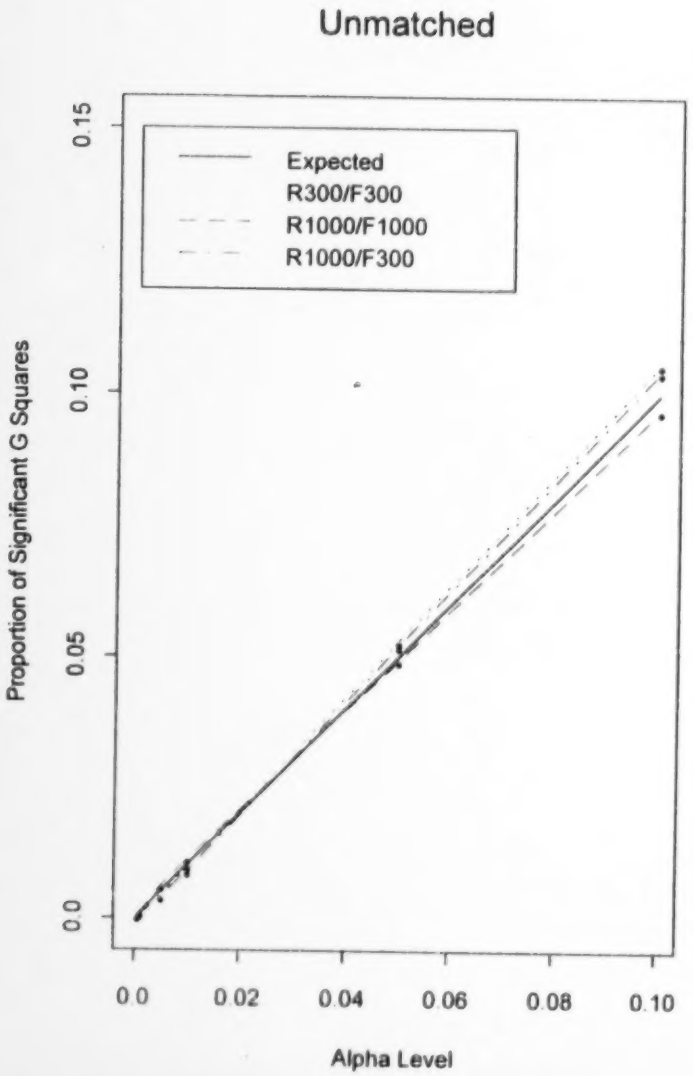
Figure 1a. Proportion of Significant G^2s for the Unmatched Ability

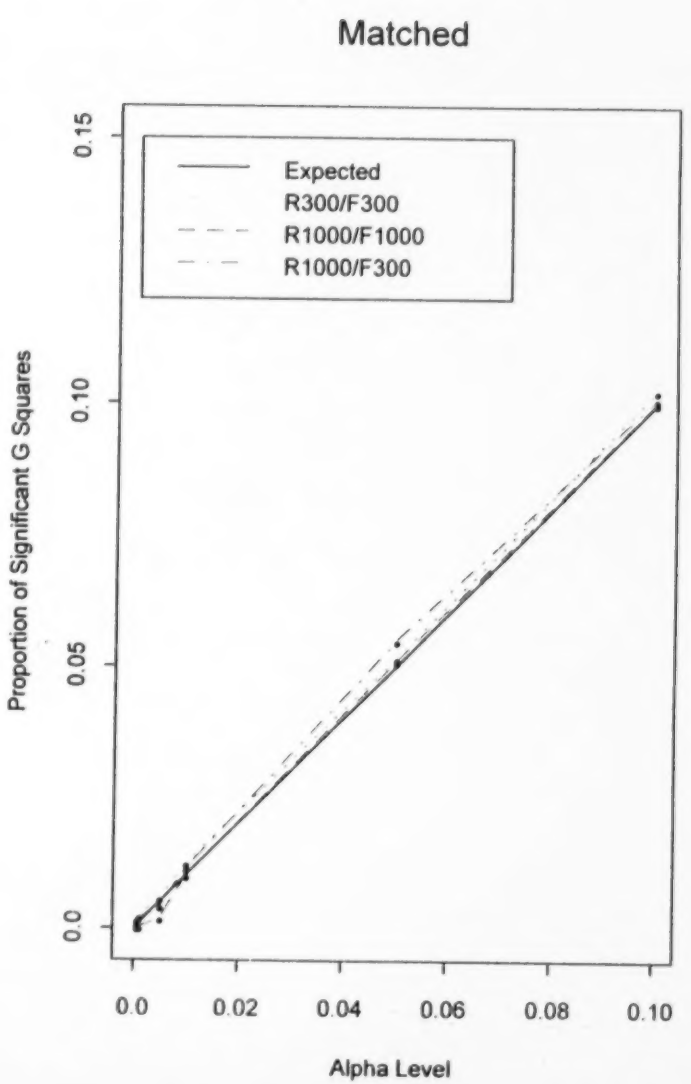
Figure 1b. Proportion of Significant G^2s for the Matched Ability

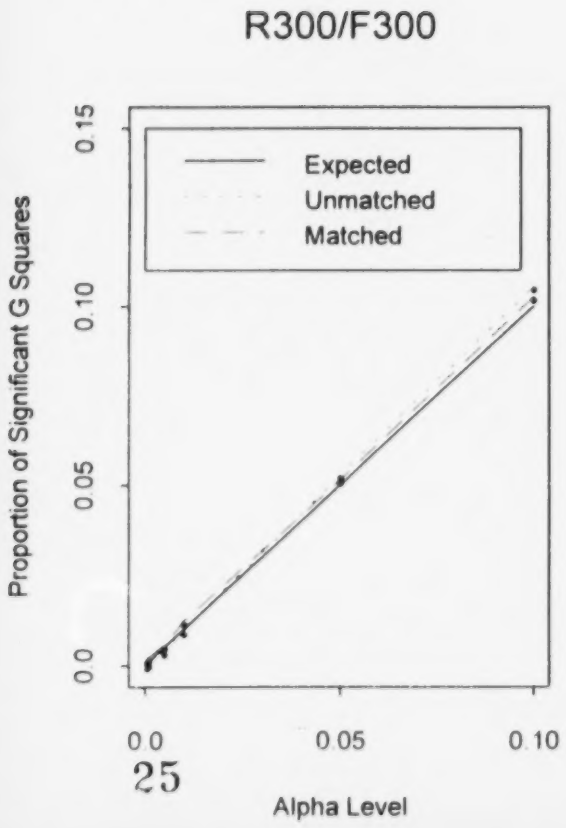
Figure 2a. Proportion of Significant G^2s for the R300/F300

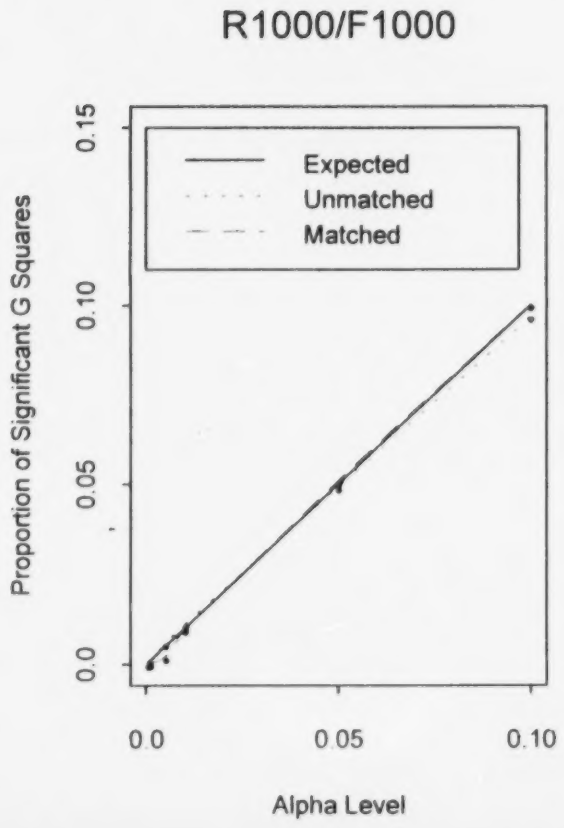
Figure 2b. Proportion of Significant G^2s for the R1000/F1000

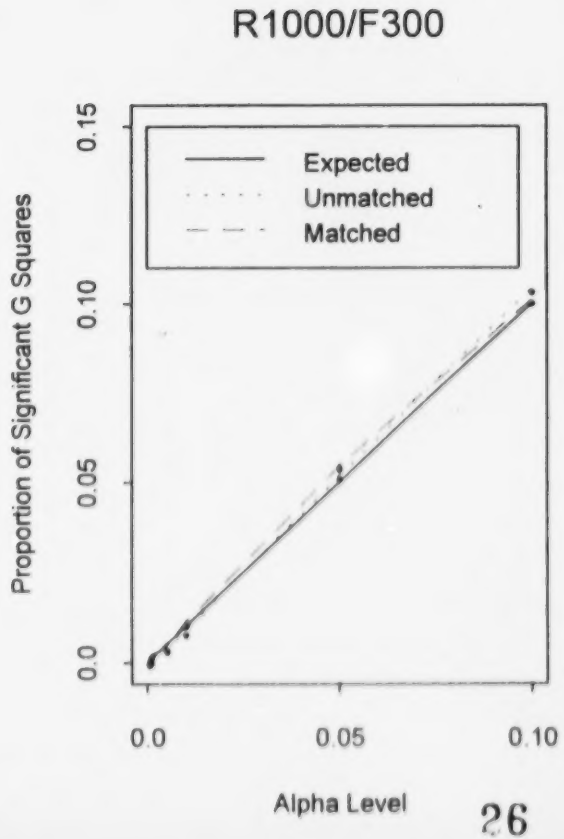
Figure 2c. Proportion of Significant G^2s for the R1000/F300


Unmatched
Expected
R300/F300
R1000/F1000
R1000/F300
Proportion of Significant G Squares
0.0
0.05
0.10
0.15
0.0
0.02
0.04
0.06
0.08
0.10
Alpha Level


Matched
Expected
R300/F300
R1000/F1000
R1000/F300
Proportion of Significant G Squares
0.0
0.05
0.10
0.15
0.0
0.02
0.04
0.06
0.08
0.10
Alpha Level


R300/F300
Proportion of Significant G Squares
0.0
0.05
0.10
0.15
Expected
Unmatched
Matched
0.0
0.05
0.10
Alpha Level


R1000/F1000
Proportion of Significant G Squares
0.0
0.05
0.10
0.15
Expected
Unmatched
Matched
0.0
0.05
0.10
Alpha Level


R1000/F300
Proportion of Significant G Squares
0.0
0.05
0.10
0.15
Expected
Unmatched
Matched
0.0
0.05
0.10
Alpha Level